MS&E 213 / CS 269O
Chapter 3 - Convexity*

By Aaron Sidford (sidford@stanford.edu)

November 11, 2019

In the last chapter we saw that we can compute an ϵ-critical point at rate independent of dimension given a gradient oracle for a smooth function. We obtained this result by showing that if $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth, then it is the case that

$$f(y) \leq f(x) + \bigtriangledown f(x)^\top (y-x) + \frac{L}{2}\|y-x\|_2^2 \tag{0.1}$$

for all $x, y \in \mathbb{R}^n$. Consequently

$$f\left(x - \frac{1}{L} \bigtriangledown f(x)\right) \leq f(x) - \frac{1}{2L}\| \bigtriangledown f(x)\|_2^2$$

and therefore, repeated gradient descent steps eventually find a point with small gradient (since otherwise too much function progress is made). While this works well to compute a critical point, it doesn't yield any global optimality guarantees.

Here we show how to add assumptions so that we can prove gradient descent doesn't just compute an ϵ-critical point, but rather it achieves an ϵ-optimal point as well. First, we motivate the assumption we make (namely *(strong) convexity*), then we prove several equivalent definitions, and then we use it to analyze gradient descent.

1 Assumptions for Proving Global Optimality

So how do we turn gradient descent from an algorithm that computes ϵ-critical points to an algorithm that computes ϵ-optimal points. It seems like we need to make another assumption. Below we discuss a few natural assumptions to make.

1.1 Assumption #1 - Lower Bound Hessian

One way that we proved the upper bound was by assuming that f was twice differentiable and that $z^\top \bigtriangledown^2 f(x) z \leq L\|z\|_2^2$ for all $x, z \in \mathbb{R}^n$. This implied that for arbitrary $x, y \in \mathbb{R}^n$ and $x_\alpha = x + \alpha(y-x)$ for all $\alpha \in [0,1]$ the following formula held

$$f(y) = f(x) + \bigtriangledown f(x)^\top (y-x) + \int_0^1 \int_0^t (y-x)^\top \bigtriangledown^2 f(x_\alpha)(y-x) d\alpha dt\,.$$

[*] These notes are a work in progress. They are not necessarily a subset or superset of the in-class material and there may also be occasional *TODO* comments which demarcate material I am thinking of adding in the future. These notes will converge to a superset of the class material that is TODO-free. Your feedback is welcome and highly encouraged. If anything is unclear, you find a bug or typo, or if you would find it particularly helpful for anything to be expanded upon, please do not hesitate to post a question on the discussion board or contact me directly at sidford@stanford.edu.

This implied (0.1) and let us show that a gradient descent step made progress proportional to the norm of the gradient. Unfortunately, this did not let us achieve global optimality as we had no way to relate the progress of a gradient descent step, i.e. the norm of the gradient, to the distance of the given point from optimality.

One natural way to fix this is to assume a lower bound on $z^\top \bigtriangledown^2 f(x)z$ namely, we could assume that for some $\mu \geq 0$ that $z^\top \bigtriangledown^2 f(x)z \geq \mu\|z\|_2^2$. Integrating this would imply that

$$f(y) \geq f(x) + \bigtriangledown f(x)^\top (y-x) + \frac{\mu}{2}\|x-y\|_2^2$$

for all $x, y \in \mathbb{R}^n$ and thus possibly allow us to relate the norm of the gradient to our current function error.

1.2 Assumption #2 - Lower Bound on Taylor Expansion

Another natural way to achieve global guarantees on the function progress would be to leverage the fact that we have already proven that gradient descent on a smooth function allows us to compute ϵ-critical points, that is points where the norm of the gradient is sufficiently small. More precisely, we could assume that if $\bigtriangledown f(x) = 0$ for some $x \in \mathbb{R}^n$ then x is a minimizer of f. To try to make this more quantitative, we note that assuming

$$f(y) \geq f(x) + \bigtriangledown f(x)^\top (y-x) + \frac{\mu}{2}\|x-y\|_2^2$$

for all $x, y \in \mathbb{R}^n$ would allow us to conclude that if $\bigtriangledown f(x) = 0$ then x is the unique global minimum of f. Thus we could just try assuming the above formula directly.

1.3 Assumption #3 - Seeing the Minima

One more way we could fix our assumptions is to take a closer look at why exactly gradient descent might get stuck. The issue with gradient descent is that for any point if we look in the direction of the minimizer of f it might be the case that the function increase rather than decreases. Thus it might be the case the locally, moving in the direction of the minimum doesn't help. More broadly, this issue is that the function might lie above the line between two points. We could fix this by assume that

$$f(t \cdot y + (1-t) \cdot x) \leq t \cdot f(y) + (1-t) \cdot f(x)$$

for all $x, y \in \mathbb{R}^n$ and $t \in [0, 1]$, meaning that the value of the function always lies underneath the line between two points. We could even make this stronger and say that for some $\mu \geq 0$ the function lies underneath the quadratic between the two points, i.e.

$$f(t \cdot y + (1-t) \cdot x) \leq t \cdot f(y) + (1-t) \cdot f(x) - \frac{\mu}{2} \cdot t \cdot (1-t) \cdot \|x-y\|_2^2 .$$

2 Convexity

It turns out that each of the possible assumptions discussed in the previous section are equivalent to a notion known as μ-strong convexity. Here we formally define μ-strong convexity and proof these equivalences.

Definition 1 ((μ-strong) convexity)**.** We say a function $f : \mathbb{R}^n \to \mathbb{R}$ is *μ-strongly convex* for $\mu \geq 0$ if and only if for all $x, y \in \mathbb{R}^n$ and $t \in [0, 1]$ we have that

$$f(t \cdot y + (1-t) \cdot x) \leq t \cdot f(y) + (1-t) \cdot f(x) - \frac{\mu}{2} \cdot t \cdot (1-t) \cdot \|x-y\|_2^2 . \tag{2.1}$$

We say that f is *convex* if this holds for $\mu = 0$.

In the remainder of this section we show that this is formally equivalent to the assumptions presented in Section 1.

Lemma 2 (Convexity of Differentiable Functions)**.** *A differentiable function f is μ-strongly convex for $\mu \geq 0$ if and only if*

$$f(y) \geq f(x) + \triangledown f(x)^\top (y-x) + \frac{\mu}{2}\|y-x\|_2^2 \text{ for all } x, y \in \mathbb{R}^n\,. \tag{2.2}$$

Proof. First suppose (2.2) holds. The for all all $x, y \in \mathbb{R}^n$ and $t \in [0,1]$ it is the case that if $x_t = x + t(y-x)$ then

$$f(y) \geq f(x_t) + \triangledown f(x_t)^\top (y - x_t) + \frac{\mu}{2}\|y - x_t\|_2^2 = f(x_t) + (1-t)\cdot \triangledown f(x_t)^\top (y-x) + \frac{\mu}{2}(1-t)^2\|y-x\|_2^2$$

and

$$f(x) \geq f(x_t) + \triangledown f(x_t)^\top (x - x_t) + \frac{\mu}{2}\|x - x_t\|_2^2 = f(x_t) - t\cdot \triangledown f(x_t)^\top (y-x) + \frac{\mu}{2}t^2\|y-x\|_2^2\,.$$

Since $t(1-t)^2 + t^2(1-t) = t(1-t)$ adding a t multiple of the first equation to the $1-t$ multiple of the second equation then yields (2.1).

On the other hand, suppose that f is μ-strongly convex. Let $x, y \in \mathbb{R}^n$ be arbitrary and let $x_t = x + t(y-x)$ then

$$f(y) \geq \frac{f(x_t) - (1-t)\cdot f(x) + \frac{\mu}{2}\cdot t\cdot(1-t)\cdot\|y-x\|_2^2}{t} = f(x) + \frac{\mu}{2}\cdot(1-t)\cdot\|y-x\|_2^2 + \frac{f(x_t)-f(x)}{t}$$

by the definition of strong convexity. However, as we have already shown that for $g(t) = f(x_t)$ it is the case that $g'(t) = \triangledown f(x_t)^\top (y-x)$ we see that taking the limit of the above as $t \to 0$ yields the desired result. □

Lemma 3 (Convexity of Twice Differentiable Functions)**.** *A twice differentiable function f is μ-strongly convex for $\mu \geq 0$ if and only if*

$$z^\top \triangledown^2 f(x) z \geq \mu\|z\|_2^2 \text{ for all } x, y \in \mathbb{R}^n \tag{2.3}$$

Proof. First suppose 2.3 holds. Then for all $x, y \in \mathbb{R}^n$ and $t \in [0,1]$ it is the case that if $x_t = x + t(y-x)$ then

$$\begin{aligned} f(y) &= f(x) + \triangledown f(x)^\top (y-x) + \int_0^1 \int_0^t (y-x)^\top \triangledown^2 f(x_\alpha)(y-x) d\alpha dt \\ &\geq f(x) + \triangledown f(x)^\top (y-x) + \int_0^1 \int_0^t \mu\|y-x\|_2^2 d\alpha dt \\ &\geq f(x) + \triangledown f(x)^\top (y-x) + \frac{\mu}{2}\|y-x\|_2^2 \end{aligned}$$

and the result follows from Lemma 2.

On the other hand suppose that f is μ-strongly convex and let $x, z \in \mathbb{R}^n$ be arbitrary. Define $x_t = x + tz$ for all $t \in \mathbb{R}$ and let $g(t) = f(x_t)$. We have $g'(t) = \triangledown f(x_t)^\top z$ and $g''(t) = z^\top \triangledown^2 f(x_t) z$. We have that

$$g''(0) = \lim_{t\to 0}\frac{g'(t) - g'(0)}{t} = \lim_{t\to 0}\frac{(\triangledown f(x_t) - \triangledown f(x))^2 z}{t} = \lim_{t\to 0}\frac{(\triangledown f(x_t) - \triangledown f(x))^\top (x_t - x)}{t^2}\,.$$

However, by Lemma 2 we know that

$$f(x_t) \geq f(x) + \triangledown f(x)^\top (x_t - x) + \frac{\mu}{2}\|x_t - x\|_2^2$$

and

$$f(x) \geq f(x_t) + \triangledown f(x_t)^\top (x - x_t) + \frac{\mu}{2}\|x_t - x\|_2^2\,.$$

Adding these and using that $x_t - x = tz$ we have

$$(\triangledown f(x_t) - \triangledown f(x))^\top (x_t - x) \geq \mu t^2 \cdot \|z\|_2^2$$

yielding the desired result. □

3 Bound on Distance to Optimum

Now that we have established the definition of convexity, we wish to use it to show that gradient descent converges to optimal points of a convex function. We have already shown that a gradient descent step from an arbitrary point decreases the function by an amount proportional to the norm of the gradient. What we want to use convexity to show is that this progress is sufficiently large relative to the current points distance to optimum.

In this section, we show how we can use the assumptions of smoothness and strong convexity to relate various measures of optimality, or difference of a point from optimum. There are three such measures of optimality that we consider. The first is one we have talked about the most, and that is *optimality* or *function error*, for a point $x \in \mathbb{R}^n$ this is just $f(x) - f_*$. The second one is simply the distance of a point to an optimum point. For a point x and minimizer x_* this is just $\|x - x_*\|_2^2$. We occasionally may refer to this as *residual error*. From here on we also let $X_*(f)$ denote the set of minimizers of f and may occasionally consider $\min_{x_* \in X_*} \|x - x_*\|_2$ which we will call the distance to the minimizing set, since this is precisely what it is.

The last measure of optimality we will occasionally consider for a point x is simply the norm of the gradient $\| \triangledown f(x)\|_2$. As we have seen, this is difficult to relate to the other two measures in general, since for non-convex functions we may have $\| \triangledown f(x)\|_2 = 0$ and nevertheless the point is optimal; however we will show that with convexity, this cannot happen.

We begin by proving bounds between these three measures using smoothness. We prove the lemma below simply by using that if $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth then for all $x, y \in \mathbb{R}^n$ we have

$$f(y) \leq f(x) + \triangledown f(x)^\top (y - x) + \frac{L}{2}\|y - x\|_2^2$$

and then look at this inequality when we set $y = x_*$, $x = x$, $y = x$ and $x = x_*$.

Lemma 4. *If $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth then for all $x_* \in X_*(f)$ we have that $\triangledown f(x_*) = 0$ and for all $x \in \mathbb{R}^n$ it is the case that*

$$\frac{1}{2L}\| \triangledown f(x)\|_2^2 \leq f(x) - f(x_*) \leq \frac{L}{2} \cdot \|x - x_*\|_2^2\,.$$

Proof. First note that $y = x - \frac{1}{L} \triangledown f(x)$ has the property that $f(y) \leq f(x) - \frac{1}{2L}\| \triangledown f(x)\|_2^2$. Considering if x was x_* we see that if $\| \triangledown f(x)\|_2$ was not equal to 0 then y would have strictly smaller value violating the optimality of x_*. Furthermore, this implies that

$$f_* \leq f(y) \leq f(x) - \frac{1}{2L}\| \triangledown f(x)\|_2^2$$

giving the right hand side of the desired identity. The right hand side follows from the fact that $f(x) \leq f(x_*) + \triangledown f(x_*)^\top (x - x_*) + \frac{L}{2}\|x - x_*\|_2^2$ by smoothness and that $\triangledown f(x_*) = \vec{0}$ as we have shown. □

With this in hand we now provide analogous bounds for μ-strongly convex functions. The proof is quite similar to the proof of the above lemma. We simply use that if $f : \mathbb{R}^n \to \mathbb{R}$ is μ-strongly convex then for all $x, y \in \mathbb{R}^n$ we have

$$f(y) \geq f(x) + \triangledown f(x)^\top (y - x) + \frac{\mu}{2}\|y - x\|_2^2$$

and then look at this inequality when we set $y = x_*$ and $x = x$ and $y = x$ and $x = x_*$. As one would expect, this lets us prove the same sort of lemma as above, with the direction of the inequalities reversed.

Lemma 5. *If $f : \mathbb{R}^n \to \mathbb{R}$ is smooth and μ-strongly convex for $\mu > 0$ then for $x_* \in X_*(f)$ we have*

$$\frac{1}{2\mu}\| \bigtriangledown f(x)\|_2^2 \geq f(x) - f(x_*) \geq \frac{\mu}{2} \cdot \|x - x_*\|_2^2 .$$

Proof. First we note that since f is smooth we have $\bigtriangledown f(x_*) = 0$ and therefore

$$f(x) \geq f(x_*) + \bigtriangledown f(x_*)^\top (x - x_*) + \frac{\mu}{2}\|x - x_*\|_2^2$$

gives the desired bounds on the right hand side. Next we note that

$$f(x_*) \geq \min_y f(x) + \bigtriangledown f(x)^\top (y - x) + \frac{\mu}{2}\|y - x\|_2^2 \geq f(x) - \frac{1}{2\mu}\| \bigtriangledown f(x)\|_2^2$$

where this follows from the analogous fact in the gradient descent proof that the minimizer of the quadratic is $y = x - \frac{1}{\mu} \bigtriangledown f(x)$. □

4 Gradient Descent Strongly Convex Case

We now have everything to analyze gradient descent for strongly convex functions. Note that Lemma 5 let us lower bound the norm of the gradient at a point by the function error at that point.

Theorem 6. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a L-smooth μ-strongly convex function for $\mu > 0$. Then for $x_0 \in \mathbb{R}^n$ let $x_{k+1} = x_k - \frac{1}{L} \bigtriangledown f(x_k)$ for all $k \geq 0$. Then we have*

$$f(x_k) - f_* \leq \left(1 - \frac{\mu}{L}\right)^k [f(x_0) - f_*]$$

and consequently we can compute an ϵ-optimal point with $\lceil \frac{L}{\mu} \log(\frac{f(x_0) - f_*}{\epsilon}) \rceil$ calls to a gradient oracle.

Proof. As we have seen $f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}\|\bigtriangledown f(x_k)\|_2^2$. However, from Lemma 5 we have that $\|\bigtriangledown f(x_k)\|_2^2 \geq 2\mu\,[f(x_k) - f_*]$. Consequently

$$f(x_{k+1}) - f_* \leq f(x_k) - f_* - \frac{\mu}{L}\,[f(x_k) - f_*] = \left(1 - \frac{\mu}{L}\right)[f(x_k) - f_*]$$

applying this repeatedly uses the bound on $f(x_k) - f_*$ and using that $1 + x \leq e^x$ for all $x \in \mathbb{R}$ and picking $k = \lceil \frac{L}{\mu} \log(\frac{f(x_0) - f_*}{\epsilon}) \rceil$ then yields the bound on the number of gradient oracle calls. □

5 Gradient Descent Non-strongly Convex Case

Here we show how to analyze gradient descent in the case when f is not strongly convex and just convex. To do this we need a new bound on the norm of the gradient and for this we prove the following.

Lemma 7. *If $f \in \mathbb{R}^n$ is differentiable and convex then for all x we have that*

$$f(x) - f_* \leq \| \bigtriangledown f(x)\|_2 \cdot \min_{x_* \in X_*} \|x - x_*\|_2 .$$

Proof. By convexity we have that

$$f_* = \max_{x_* \in X_*} f(x_*) \geq \max_{x_* \in X_*} f(x) + \triangledown f(x)^\top (x_* - x) \geq f(x) - \| \triangledown f(x) \| \cdot \min_{x_* \in X_*} \|x - x_*\|_2$$

where in the last step we used that $\triangledown f(x)^\top (y - x) \geq - \left| \triangledown f(x)^\top (y - x) \right|$ and $\left| \triangledown f(x)^\top (x_* - x) \right| \leq \| \triangledown f(x) \|_2 \cdot \|x_* - x\|_2$ by Cauchy Schwarz. □

Using this, we have everything we need to analyze gradient descent.

Theorem 8 (Gradient Descent)**.** *Let f be a L-smooth convex function and starting from some x_0 let $x_{k+1} = x_k - \frac{1}{L} \triangledown f(x_k)$. Then if $D = \max_{y \,:\, f(y) \leq f(x_0)} \min_{x_* \in X_*} \|y - x_*\|_2$ we have that*

$$f(x_{k+1}) - \min_x f(x) \leq \frac{2 \cdot L \cdot D^2}{k + 4} \,.$$

and consequently we can compute an ϵ-optimal point with $\lceil 2 \cdot L \cdot D^2 / \epsilon \rceil$ calls to a gradient oracle.

Proof. Let $\epsilon_k \stackrel{\text{def}}{=} f(x_k) - f_*$. We have already argued by smoothness that $f(x_{k+1}) \leq f(x_k) - \frac{1}{2L} \| \triangledown f(x_k) \|_2^2$ and therefore $\epsilon_{k+1} \leq \epsilon_k - \frac{1}{2L} \| \triangledown f(x_k) \|_2^2$. Furthermore, by Lemma 7 we have that since $f(x_{k+1}) \leq f(x_0)$ for all k it is the case that $\min_{x_* \in X_*} \|x_k - x_*\|_\infty \leq D$ and thus $\epsilon_k \leq \| \triangledown f(x_k) \|_2 \cdot D$.

Combining yields that

$$\epsilon_{k+1} \leq \epsilon_k - \frac{\epsilon_k^2}{2 \cdot L \cdot D^2}$$

and therefore

$$\frac{1}{\epsilon_{k+1}} - \frac{1}{\epsilon_k} \geq \frac{\epsilon_k - \epsilon_{k+1}}{\epsilon_k \epsilon_{k+1}} \geq \frac{\epsilon_k}{2 \cdot L \cdot D^2 \cdot \epsilon_{k+1}} \geq \frac{1}{2 \cdot L \cdot D^2} \,.$$

Since $\epsilon_0 \leq \frac{L}{2} D^2$ by Lemma 4 we have that c

$$\frac{1}{\epsilon_k} \geq \frac{1}{\epsilon_0} + \frac{k}{2 \cdot L \cdot D^2} \geq \frac{k + 4}{2 \cdot L \cdot D^2} \,.$$

□

6 Summary

In these notes we bounded the performance for gradient descent on smooth convex functions. In all of the settings we considered the algorithm, gradient descent, remained unchanged. It was only the analysis that changed. Thus we can combine the bounds to get a clean statement for the performance of gradient descent. To simplify our statements we use the following helper lemma.

Lemma 9. *If f is a L-smooth convex function and $x_* \in X_*(f)$ be then if $y = x - \eta \triangledown f(x)$ for $\eta \in [0, \frac{1}{L}]$ the following holds*

$$\|y - x_*\|_2^2 \leq \|x - x_*\|_2^2 \,.$$

Proof. By definitions we have that

$$\begin{aligned} \|y - x_*\|_2^2 &= \|x - x_* - \eta \triangledown f(x_k)\|_2^2 \\ &= \|x - x_*\|_2^2 + 2\eta \triangledown f(x)^\top (x_* - x) + \eta^2 \| \triangledown f(x) \|_2^2 \,. \end{aligned}$$

Now $f(x_*) \geq f(x) + \bigtriangledown f(x)^\top (x_* - x)$ by convexity and $\| \bigtriangledown f(x)\|_2^2 \leq 2L \cdot [f(x) - f(x_*)]$ by smoothness. Consequently

$$\begin{aligned} \|y - x_*\|_2^2 &\leq \|x - x_*\|_2^2 - 2\eta\,[f(x) - f_*] + 2\eta^2 L\,[f(x) - f_*] \\ &= \|x - x_*\|_2^2 - 2\eta(1 - \eta L) \cdot [f(x) - f_*] \ . \end{aligned}$$

Since $f(x) - f_* \geq 0$ by definition of f_* and $2\eta(1 - \eta L) \geq 0$ by assumption on η we have that

$$-2\eta(1 - \eta L) \cdot [f(x) - f_*] \leq 0$$

□

Using this, we can summarize our results as follows.

Theorem 10. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a L-smooth μ-strongly convex function for $\mu \geq 0$. Let $x_0 \in \mathbb{R}^n$ and $x_* \in X_*(f)$ be arbitrary and let $x_{k+1} = x_k - \frac{1}{L} \bigtriangledown f(x_k)$ for all $k \geq 0$. Then*

$$f(x_k) - f_* \leq \min \left\{ \left(1 - \frac{\mu}{L}\right)^k [f(x_0) - f_*] \ , \ \frac{2L \cdot \|x_0 - x_*\|_2^2}{k + 4} \right\} .$$

Consequently we can compute an ϵ-optimal point with $O(\lceil \min\{\frac{L}{\mu} \log(\frac{f(x_0) - f_*}{\epsilon}), \frac{L\|x_0 - x_*\|_2^2}{\epsilon}\}\rceil)$ oracle calls.

Proof. Combine the theorems in this chapter regarding gradient descent and note that the $f(x_k) - f_* \leq \left(1 - \frac{\mu}{L}\right)^k [f(x_0) - f_*]$ still holds when $\mu = 0$ as each step of gradient descent reduces the function value. □